

MARVEL & DC DATASET

Viero Hedfam Putri - Kelas A SIB GNFI 7

ELEMENTS OF DATASET

ID

Berfungsi sebagai identitas atau pengenal untuk setiap film atau serial.

Movie

Berisi judul film atau serial.

Year

Menunjukkan tahun rilis atau tahun produksi film atau serial tersebut.

Genre

Berisi genre atau kategori dari film atau serial tersebut.

RunTime

Menunjukkan durasi atau lamanya waktu yang dibutuhkan untuk menonton keseluruhan film atau episode serial.

Description

Berisi tentang deskripsi singkat mengenai isi atau cerita dari film atau serial tersebut.

IMDB_Score

Menunjukkan rating atau skor yang diberikan oleh penggunaan situs web IMDb untuk film atau serial tersebut.

DATASET

Dataset terdiri dari 1690 baris
dan 7 kolom

ID	Movie	Year	Genre	RunTime	Description	IMDB_Score
0	Eternals	-2021	Action,Adventure,Drama	0	The saga of the Eternals, a race of immortal b...	0.0
1	Loki	(2021–)	Action,Adventure,Fantasy	0	A new Marvel chapter with Loki at its center.	0.0
2	The Falcon and the Winter Soldier	-2021	Action,Adventure,Drama	50 min	Following the events of 'Avengers: Endgame,' S...	7.5
3	WandaVision	-2021	Action,Comedy,Drama	350 min	Blends the style of classic sitcoms with the M...	8.1
4	Spider-Man: No Way Home	-2021	Action,Adventure,Sci-Fi	0	A continuation of Spider-Man: Far From Home.	0.0
...
1685	DC's Legends of Tomorrow	(2016–)	Action,Adventure,Drama	42 min	Worlds lived, worlds died. Nothing will ever b...	8.5
1686	Supergirl	(2015–2021)	Action,Adventure,Drama	42 min	In the wake of Lex Luthor's return, the show f...	8.3
1687	Supergirl	(2015–2021)	Action,Adventure,Drama	42 min	Kara comes face to face with Red Daughter and ...	8.1
1688	Supergirl	(2015–2021)	Action,Adventure,Drama	42 min	Kara and Lena head to Kaznia to hunt down Lex....	7.4
1689	Supergirl	(2015–2021)	Action,Adventure,Drama	42 min	Supergirl must deal with the destructive after...	7.5

1690 rows × 7 columns

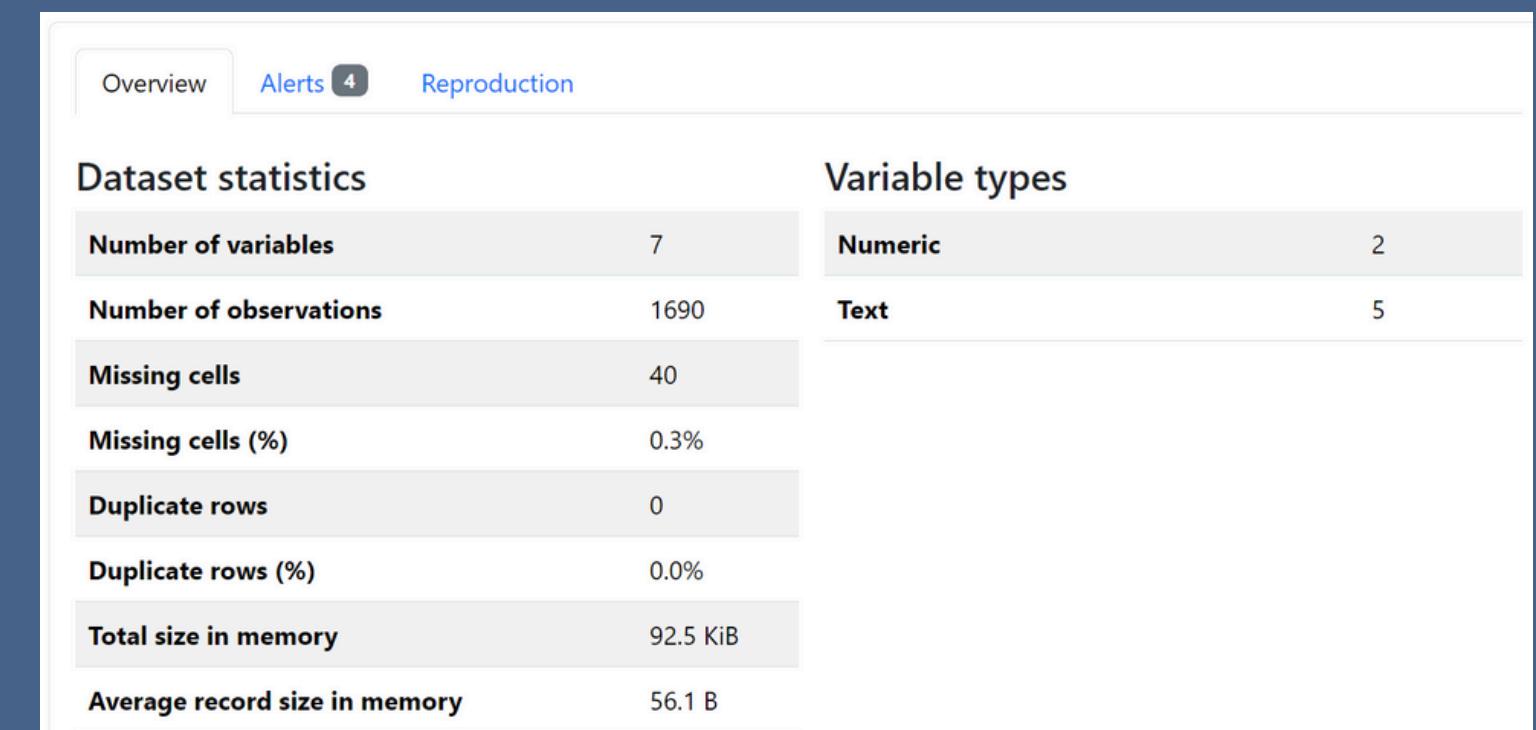
```

1 import pandas as pd
2 import numpy as np
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5 from ydata_profiling import ProfileReport
6
7 data=pd.read_csv("Marvel Vs DC NEW.csv")
8 report=ProfileReport(data, title="Marvel vs DC")
9 report

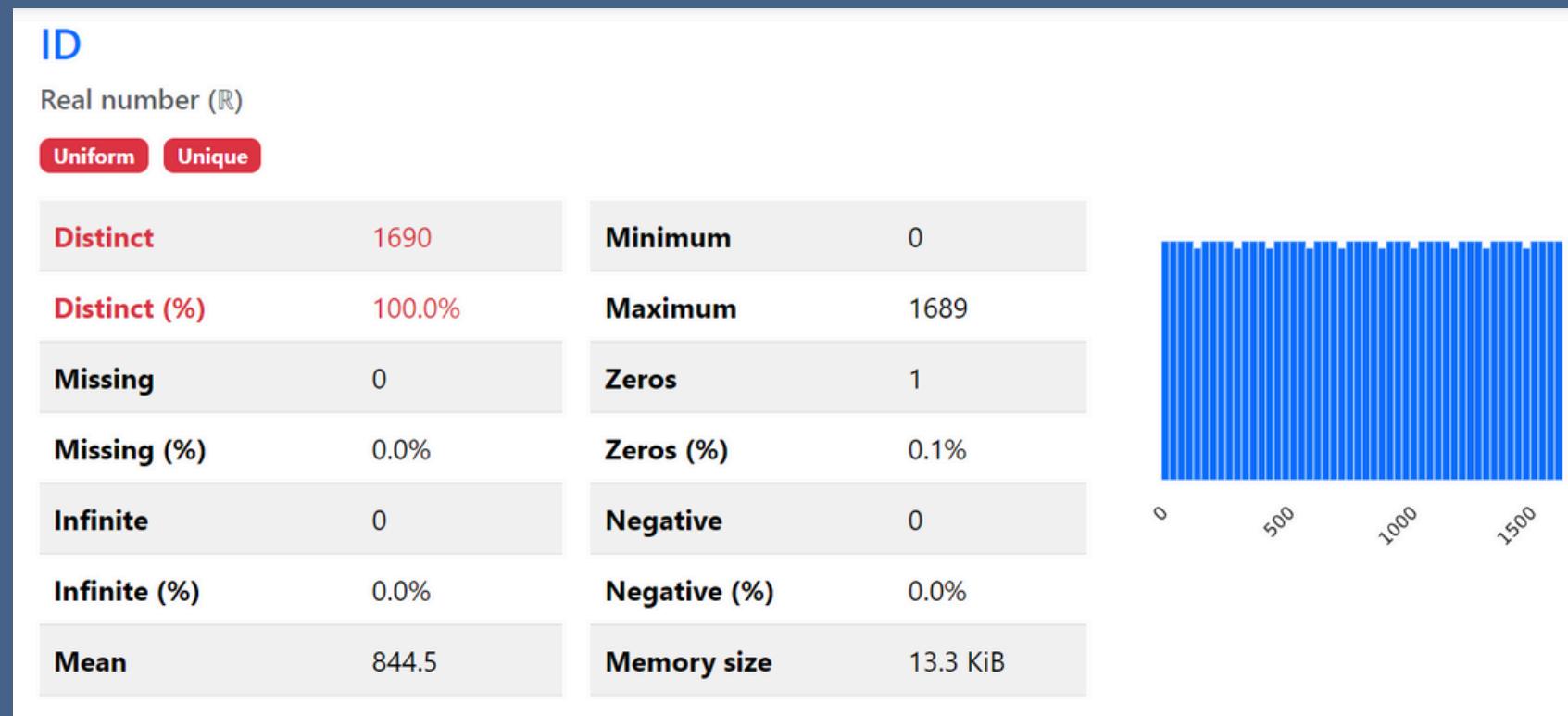
```

- Dataset ini memiliki 2 variabel numerik (angka) dan 5 variabel teks.
- Terdapat 40 sel data yang kosong atau hilang, yang setara dengan 0,3% dari total data.
- Tidak ada baris data yang duplikat atau sama persis.
- Dataset ini berisi informasi yang sebagian besar berupa angka dan sebagian lagi berupa teks.

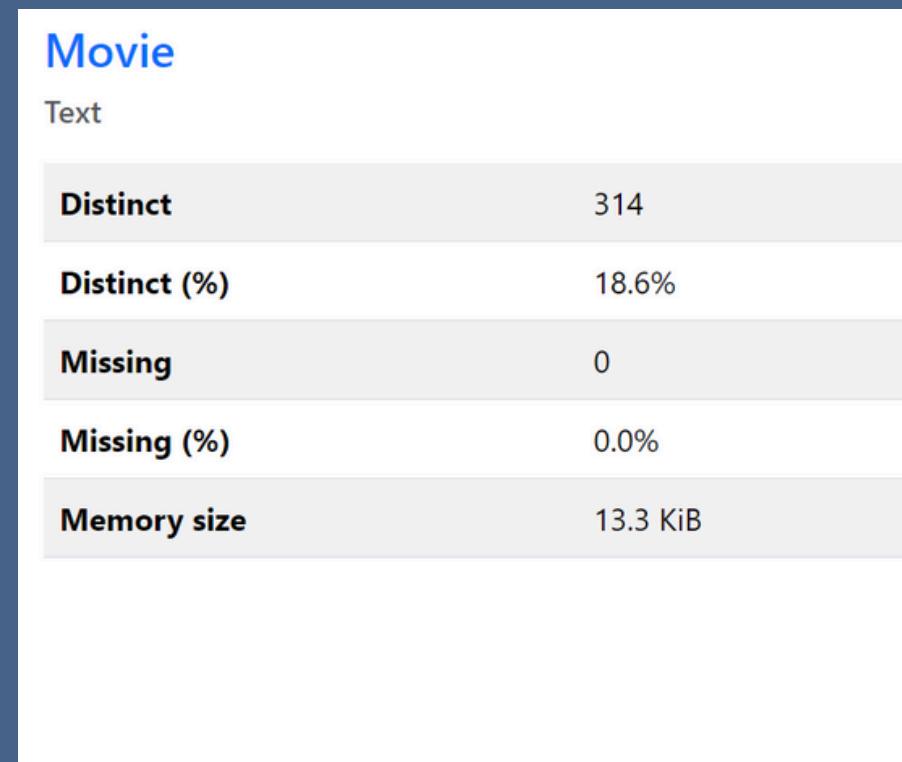
Meskipun terdapat sedikit data yang hilang, secara keseluruhan data cukup lengkap dan tidak ada duplikasi data. Ukuran data relatif kecil, sehingga mudah untuk diproses.



- Terdapat 1690 ID numerik unik, mulai dari 0 hingga 1689.
 - Distribusi ID cukup merata, seperti yang terlihat pada grafik batang.
 - Tidak ada data yang hilang atau memiliki nilai ekstrem (tak terhingga atau negatif).
 - Hanya 1 ID yang bernilai nol.
 - Mean nilai ID adalah 844.5.



- Cloud word yang menampilkan kata-kata kunci paling sering muncul dalam data film.
 - Terdapat 314 kata unik yang terkait dengan film, menunjukkan beragam topik yang dibahas.
 - Kata-kata kunci yang paling menonjol berkaitan dengan genre superhero, terutama dari universe DC Comics.
 - Data yang dianalisis cukup lengkap tanpa adanya data yang hilang.



Year	
Text	
Missing	
Distinct	147
Distinct (%)	8.9%
Missing	33
Missing (%)	2.0%
Memory size	13.3 KiB



Genre	
Text	
Distinct	90
Distinct (%)	5.3%
Missing	7
Missing (%)	0.4%
Memory size	13.3 KiB

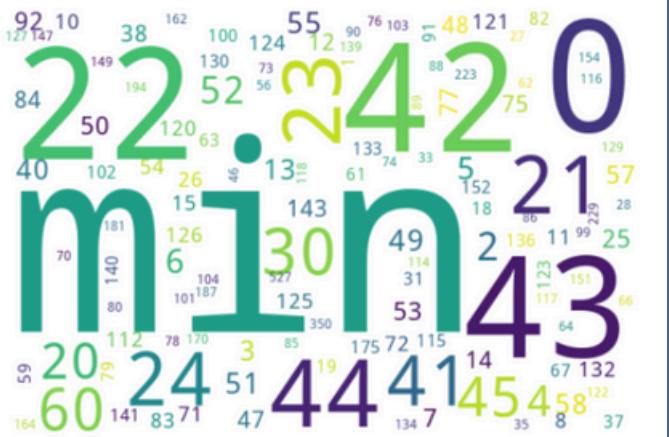


- Cloud word yang menampilkan distribusi tahun.
- Data mencakup rentang waktu yang cukup luas, dari tahun 1992 hingga 2021.
- Sekitar 2% data tidak memiliki informasi tahun.
- Distribusi tahun ini memberikan gambaran tentang periode waktu yang paling banyak terwakili dalam data.

- Cloud word yang menampilkan distribusi genre film.
- Terdapat 90 genre film yang unik, menunjukkan beragam jenis film dalam dataset.
- Sekitar 0,4% data tidak memiliki informasi genre.
- Distribusi genre ini memberikan gambaran umum tentang jenis film yang paling banyak ditemukan dalam dataset.

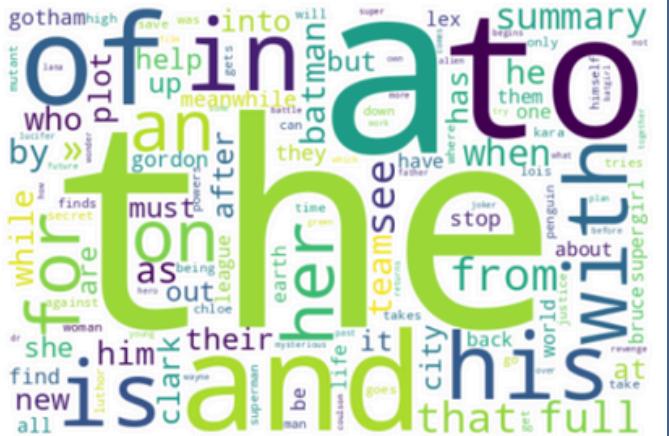
- Terdapat 126 nilai runtime yang unik, menunjukkan beragam waktu yang dibutuhkan untuk menyelesaikan proses.
- Distribusi nilai runtime tidak merata, dengan beberapa nilai yang muncul lebih sering daripada yang lainnya.
- Tidak ada data runtime yang hilang.
- Distribusi waktu eksekusi ini memberikan gambaran umum tentang performa sistem atau proses yang sedang diukur.

RunTime	
Text	
Distinct	126
Distinct (%)	7.5%
Missing	0
Missing (%)	0.0%
Memory size	13.3 KiB

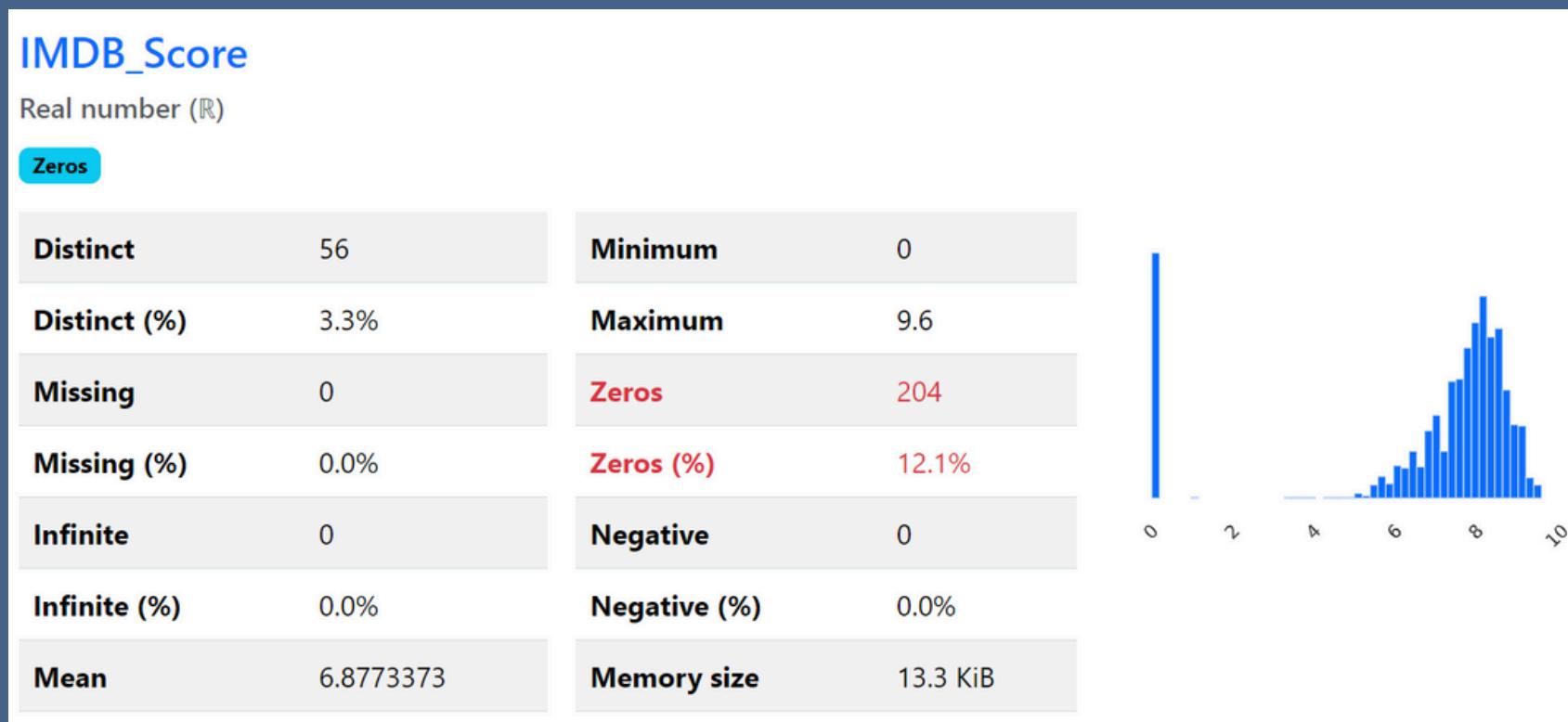


- Terdapat 1571 kata unik yang ditemukan dalam deskripsi, menunjukkan beragam kata yang digunakan.
- Tidak ada data yang hilang.

Description	
Text	
Distinct	1571
Distinct (%)	93.0%
Missing	0
Missing (%)	0.0%
Memory size	13.3 KiB



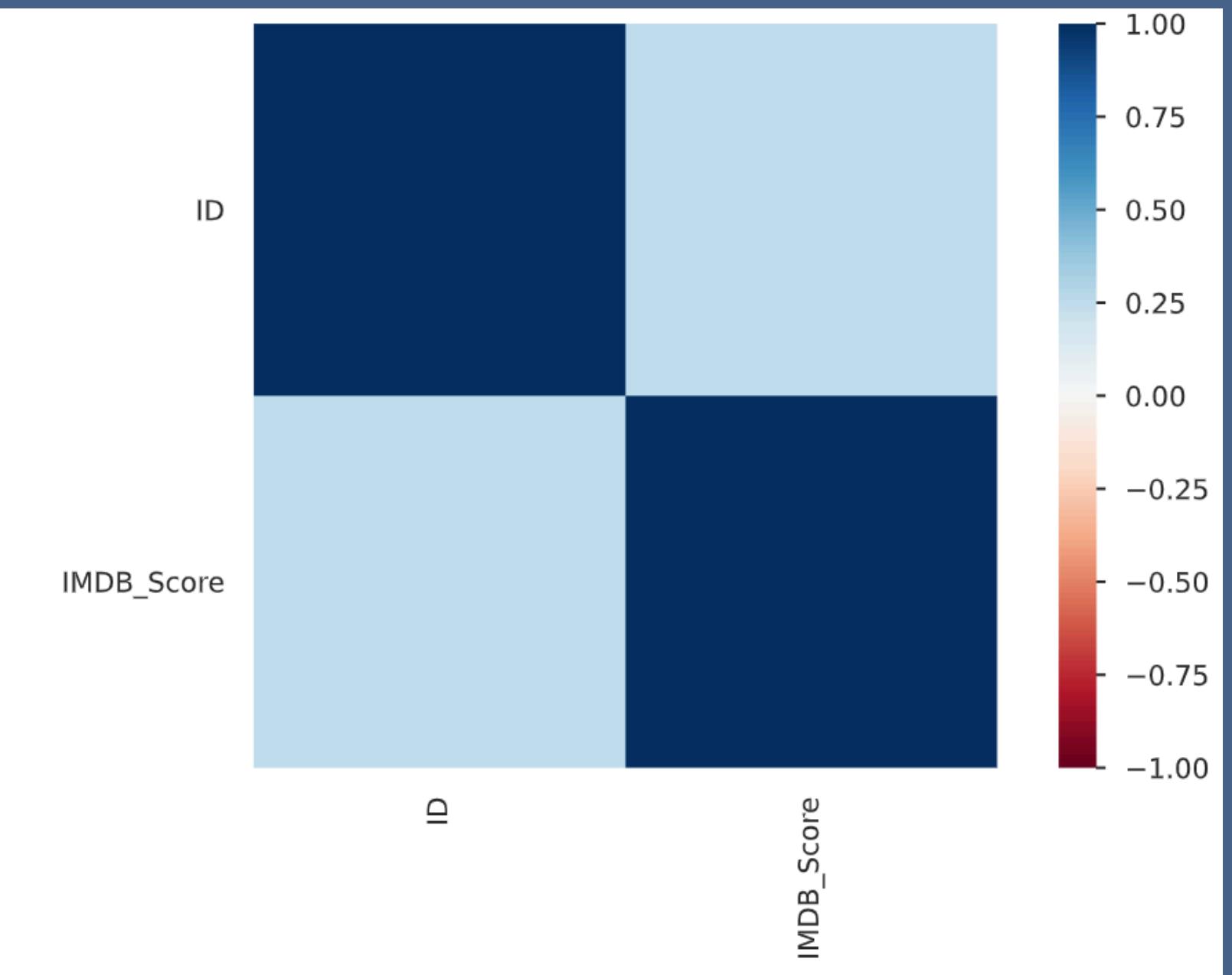
- Histogram yang menunjukkan distribusi skor IMDb.
- Terdapat 56 skor IMDb yang unik, menunjukkan beragam penilaian.
- Sebagian besar film atau serial memiliki skor di sekitar 6-8, menunjukkan penilaian yang cukup baik secara umum.
- Sebanyak 12,1% data memiliki skor 0, mungkin karena belum dinilai atau penilaian dihapus.
- Skor terendah adalah 0 dan tertinggi adalah 9.6.
- Rata-rata skor adalah 6.877.
- Distribusi skor IMDb ini memberikan gambaran umum tentang kualitas film atau serial dalam dataset.



- Sebagian besar skor berkumpul di sekitar nilai tengah, menunjukkan rata-rata penilaian yang baik.
- Terdapat pola pengelompokan skor yang tinggi, mengindikasikan adanya faktor lain yang mempengaruhi penilaian.
- Terdapat variasi yang cukup besar dalam skor, menunjukkan perbedaan kualitas dalam dataset.
- Tidak ada hubungan langsung antara ID dan skor, menunjukkan bahwa faktor lain lebih dominan dalam menentukan skor.
- Visualisasi ini memberikan gambaran umum tentang distribusi kualitas film atau serial dalam dataset berdasarkan ID.



- Analisis korelasi menunjukkan bahwa tidak ada hubungan linear yang signifikan antara ID dan Skor IMDb.
- Perubahan pada ID tidak menyebabkan perubahan yang sistematis pada Skor IMDb.
- Hasil ini menunjukkan bahwa ID film atau serial tidak dapat digunakan untuk memprediksi skor IMDb.



- Dataset memiliki kualitas yang baik dengan sedikit sekali data yang hilang.
- Hampir semua kolom memiliki 1690 data yang lengkap.
- Terdapat sedikit data hilang pada kolom "Year", "Genre", dan "Runtime".
- Dataset siap untuk dianalisis lebih lanjut tanpa masalah signifikan terkait data yang hilang.



DATA CLEANING

```
1 data=data.dropna()  
2 data=data.drop("ID", axis=1)  
3 data=data.drop("Description", axis=1)  
4 data
```

	Movie	Year	Genre	RunTime	IMDB_Score	
0	Eternals	-2021	Action, Adventure, Drama	0	0.0	  
1	Loki	(2021–)	Action, Adventure, Fantasy	0	0.0	
2	The Falcon and the Winter Soldier	-2021	Action, Adventure, Drama	50 min	7.5	
3	WandaVision	-2021	Action, Comedy, Drama	350 min	8.1	
4	Spider-Man: No Way Home	-2021	Action, Adventure, Sci-Fi	0	0.0	
...
1685	DC's Legends of Tomorrow	(2016–)	Action, Adventure, Drama	42 min	8.5	
1686	Supergirl	(2015–2021)	Action, Adventure, Drama	42 min	8.3	
1687	Supergirl	(2015–2021)	Action, Adventure, Drama	42 min	8.1	
1688	Supergirl	(2015–2021)	Action, Adventure, Drama	42 min	7.4	
1689	Supergirl	(2015–2021)	Action, Adventure, Drama	42 min	7.5	

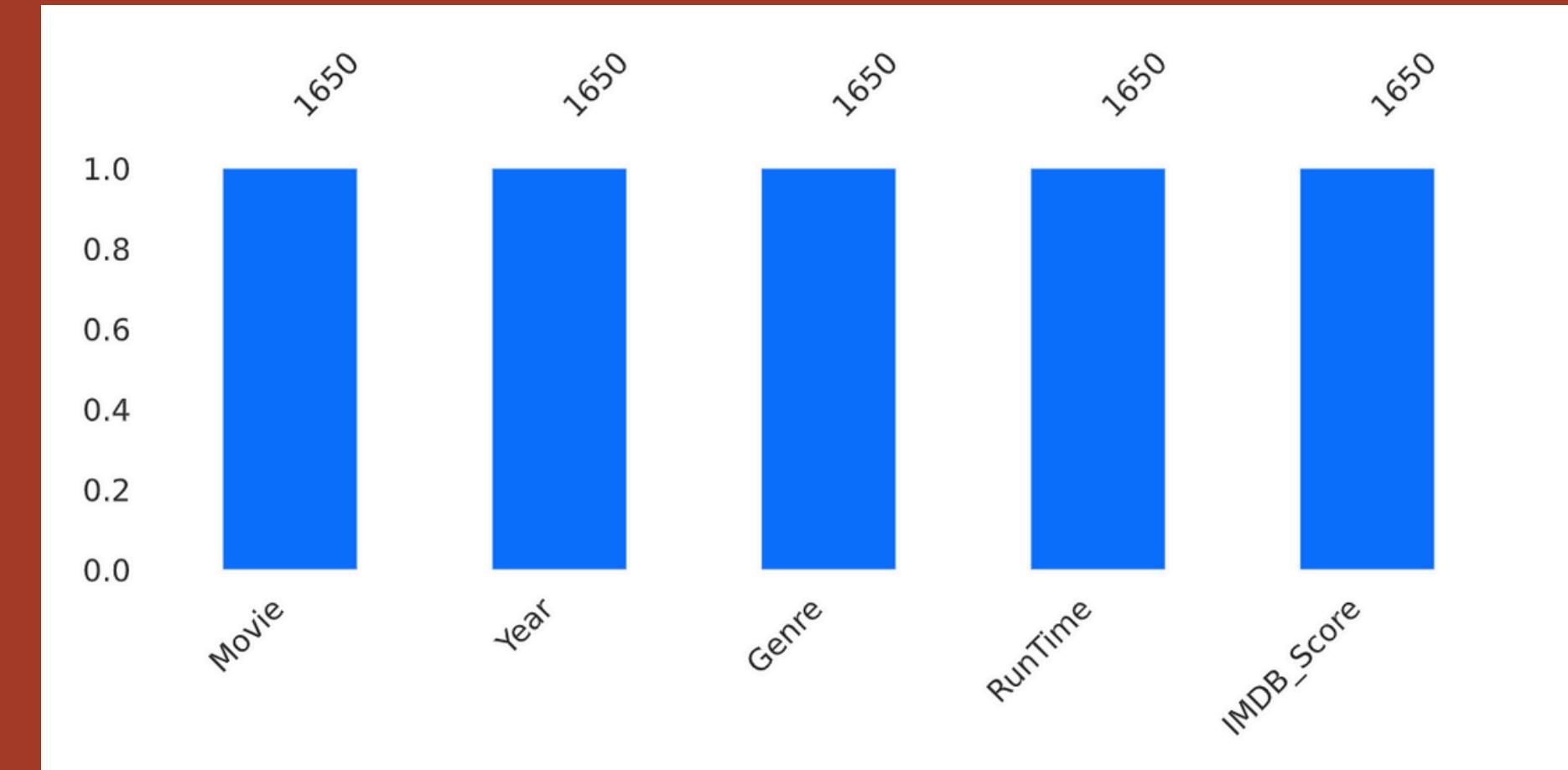
1650 rows × 5 columns

- Melakukan pembersihan data awal pada dataset film.
- Kolom ID dihapus karena dianggap tidak relevan untuk analisis selanjutnya.
- Kolom Description dihapus karena data teksnya tidak cocok untuk analisis numerik yang akan dilakukan.
- Setelah pembersihan, dataset sekarang hanya berisi informasi numerik dan kategorikal yang relevan untuk analisis lebih lanjut, seperti judul film, tahun rilis, genre, durasi, dan skor IMDb.
- Beberapa kolom dihapus untuk memfokuskan pada variabel-variabel yang dianggap penting untuk menjawab pertanyaan penelitian.

HASIL DATA CLEANING

- Setiap batang pada grafik mewakili satu kolom (variabel) dalam dataset, dan tinggi batang menunjukkan jumlah data yang lengkap (tidak hilang) pada kolom tersebut.
- Setelah proses pembersihan data, semua kolom memiliki jumlah data yang sama, yaitu 1650 data. Ini mengindikasikan bahwa tidak ada lagi data yang hilang atau nilai yang kosong pada setiap kolom.

```
1 data=data.dropna()  
2 report=ProfileReport(data, title="Data Bersih")  
3 report
```



- Tabel yang menampilkan daftar film dengan duplikasi data dan frekuensinya.
- Teridentifikasi adanya duplikasi data pada beberapa judul film dalam dataset.
- Duplikasi data dapat menyebabkan bias dalam analisis jika tidak ditangani dengan benar.
- Langkah selanjutnya adalah melakukan pembersihan data untuk menghapus duplikasi dan memastikan akurasi analisis.

Duplicate rows

Most frequently occurring

	Movie	Year	Genre	RunTime	IMDB_Score	# duplicates
60	DC Super Hero Girls	(2015–2018)	Animation,Short,Action	0	0.0	74
156	She-Hulk	(2022–)	Action,Adventure,Comedy	0	0.0	11
171	Smallville	(2001–2011)	Adventure,Drama,Romance	42 min	8.6	11
236	What If...?	(2021–)	Animation,Action,Adventure	0	0.0	11
45	Batman: The Animated Series	(1992–1995)	Animation,Action,Adventure	22 min	7.8	10
151	Mutant X	(2001–2004)	Action,Adventure,Drama	43 min	6.1	9
167	Smallville	(2001–2011)	Adventure,Drama,Romance	42 min	8.2	9
198	Supergirl	(2015–2021)	Action,Adventure,Drama	42 min	7.2	9
44	Batman: The Animated Series	(1992–1995)	Animation,Action,Adventure	22 min	7.7	8
48	Batman: The Animated Series	(1992–1995)	Animation,Action,Adventure	22 min	8.1	8

PEMBERSIHAN DATA DUPLIKAT

```
1 # Menghapus duplikat berdasarkan semua kolom
2 data = data.drop_duplicates()
3
4 # Menampilkan data setelah menghapus duplikat
5 print(data)

          Movie      Year \
0        Eternals    -2021
1         Loki      (2021- )
2 The Falcon and the Winter Soldier   -2021
3           WandaVision   -2021
4 Spider-Man: No Way Home   -2021
...          ...
1670      Supergirl (2015-2021)
1681        Arrow   (2012-2020)
1682       Krypton (2018-2019)
1684     Black Lightning (2017-2021)
1685 DC's Legends of Tomorrow   (2016- )

      Genre RunTime  IMDB_Score
0 Action, Adventure, Drama      0     0.0
1 Action, Adventure, Fantasy    0     0.0
2 Action, Adventure, Drama    50 min    7.5
3 Action, Comedy, Drama  350 min    8.1
4 Action, Adventure, Sci-Fi      0     0.0
...          ...
1670 Action, Adventure, Drama  42 min    8.6
1681 Action, Adventure, Crime  42 min    8.7
1682 Action, Adventure, Drama  41 min    7.8
1684 Action, Drama, Sci-Fi    42 min    6.9
1685 Action, Adventure, Drama  42 min    8.5

[980 rows x 5 columns]
```

- Melakukan pembersihan data dengan menghapus baris-baris yang duplikat.
- Menggunakan fungsi `drop_duplicates()` untuk menghilangkan semua baris yang identik.
- Setelah pembersihan, jumlah baris data berkurang secara signifikan, menunjukkan bahwa banyak data yang duplikat telah dihapus.
- Menghapus data duplikat penting untuk memastikan akurasi analisis data. Data duplikat dapat menyebabkan bias dalam hasil analisis jika tidak dihilangkan.

KONVERSI TIPE DATA KOLOM RUNTIME

```
1 data['RunTime'] = data['RunTime'].str.replace(' min', '').astype(int)
2 data

<ipython-input-9-e0ceb9d39fe2>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
data['RunTime'] = data['RunTime'].str.replace(' min', '').astype(int)

      Movie        Year   Genre  RunTime  IMDB_Score
0     Eternals    -2021  Action,Adventure,Drama      0       0.0
1        Loki  (2021– )  Action,Adventure,Fantasy      0       0.0
2  The Falcon and the Winter Soldier    -2021  Action,Adventure,Drama     50       7.5
3    WandaVision    -2021  Action,Comedy,Drama    350       8.1
4  Spider-Man: No Way Home    -2021  Action,Adventure,Sci-Fi      0       0.0
...
1670      ...        ...
1681      ...        ...
1682      ...        ...
1684      ...        ...
1685      ...        ...

980 rows × 5 columns
```

	Movie	Year	Genre	RunTime	IMDB_Score
0	Eternals	-2021	Action,Adventure,Drama	0	0.0
1	Loki	(2021–)	Action,Adventure,Fantasy	0	0.0
2	The Falcon and the Winter Soldier	-2021	Action,Adventure,Drama	50	7.5
3	WandaVision	-2021	Action,Comedy,Drama	350	8.1
4	Spider-Man: No Way Home	-2021	Action,Adventure,Sci-Fi	0	0.0
...
1670	Supergirl	(2015–2021)	Action,Adventure,Drama	42	8.6
1681	Arrow	(2012–2020)	Action,Adventure,Crime	42	8.7
1682	Krypton	(2018–2019)	Action,Adventure,Drama	41	7.8
1684	Black Lightning	(2017–2021)	Action,Drama,Sci-Fi	42	6.9
1685	DC's Legends of Tomorrow	(2016–)	Action,Adventure,Drama	42	8.5

- Membersihkan dan mengubah tipe data kolom 'Runtime' agar sesuai untuk analisis numerik.
- Kata "min" dihapus dari setiap nilai dalam kolom Runtime karena kata ini tidak diperlukan untuk perhitungan numerik.
- Tipe data kolom 'Runtime' diubah menjadi integer agar dapat digunakan dalam operasi matematika.
- Setelah pembersihan, kolom 'Runtime' sekarang hanya berisi nilai numerik dalam satuan menit.

DETEKSI OUTLIER DENGAN IQR

```
1 def detect_outliers(data):  
2  
3     q1 = np.quantile(data, 0.25)  
4     q3 = np.quantile(data, 0.75)  
5     iqr = q3 - q1  
6     lower_bound = q1 - 1.5 * iqr  
7     upper_bound = q3 + 1.5 * iqr  
8  
9     outliers = []  
10    for i, x in enumerate(data):  
11        if x < lower_bound or x > upper_bound:  
12            outliers.append(i)  
13  
14    return outliers  
15  
16 outliers = detect_outliers(data['RunTime'])  
17 print(outliers)
```

```
[3, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 40, 41, 42, 46, 56, 57, 59, 62, 63, 67, 71, 72]
```

- Mengidentifikasi outlier pada kolom 'Runtime' menggunakan metode interquartile range (IQR).
- Hitung kuartila pertama (Q1) dan ketiga (Q3), hitung IQR, mentukan batas atas dan bawah outlier berdasarkan Q1, Q3, dan IQR, bandingkan setiap nilai dengan batas-batas tersebut untuk mengidentifikasi outlier.
- Fungsinya untuk membantu membersihkan data sebelum analisis lebih lanjut, meningkatkan akurasi model prediksi, mendeteksi adanya anomali atau kesalahan dalam data.
- Langkah Selanjutnya: Analisis lebih lanjut terhadap outlier yang teridentifikasi (misalnya, mencari penyebabnya).
- Memutuskan tindakan yang tepat untuk menangani outlier (misalnya, menghapus, mengganti, atau menyelidiki lebih lanjut).
- Outlier mewakili film dengan durasi yang sangat panjang atau pendek, yang mungkin penting untuk analisis. Menghapusnya dapat menyebabkan hilangnya informasi berharga, maka outlier tidak akan dihapus.

```
1 def detect_outliers(data):
2
3     q1 = np.quantile(data, 0.25)
4     q3 = np.quantile(data, 0.75)
5     iqr = q3 - q1
6     lower_bound = q1 - 1.5 * iqr
7     upper_bound = q3 + 1.5 * iqr
8
9     outliers = []
10    for i, x in enumerate(data):
11        if x < lower_bound or x > upper_bound:
12            outliers.append(i)
13
14    return outliers
15
16 outliers = detect_outliers(data['IMDB_Score'])
17 print(outliers)
```

```
[0, 1, 4, 5, 8, 14, 23, 29, 30, 36, 37, 38, 39, 43, 44, 47, 51, 74, 76, 86, 87, 88, 89, 110, 132, 139, 140, 147, 148, 222, 225, 226, 237, 249, 250, 251]
```

- Mengidentifikasi data yang sangat menyimpang (outlier) pada kolom 'IMDB_Score' menggunakan metode interquartile range (IQR).
- Menentukan kuartila pertama (Q1) dan ketiga (Q3) dari data untuk mengetahui sebaran data, menghitung selisih antara Q3 dan Q1 untuk mendapatkan rentang interkuartil, menghitung batas atas dan bawah outlier berdasarkan Q1, Q3, dan IQR. Nilai yang berada di luar batas ini dianggap sebagai outlier.
- Identifikasi Outlier: Membandingkan setiap nilai data dengan batas-batas yang telah ditentukan. Jika nilai tersebut berada di luar batas, maka nilai tersebut dianggap sebagai outlier.
- Outlier dapat mewakili film dengan rating yang sangat tinggi atau rendah, yang mungkin penting untuk dianalisis. Menghapusnya dapat menyebabkan hilangnya informasi berharga, maka outlier tidak akan dihapus.

NORMALISASI DATA MENGGUNAKAN MIN-MAX SCALER SETELAH DETEKSI OUTLIER

```
1 from sklearn.preprocessing import MinMaxScaler
2 data = pd.DataFrame(data)
3 scaler = MinMaxScaler()
4 data['IMDB_Score_scaled'] = scaler.fit_transform(data[['IMDB_Score']])
5 print(data)
```

```
          Movie      Year \
0           Eternals   -2021
1            Loki    (2021- )
2  The Falcon and the Winter Soldier   -2021
3        WandaVision   -2021
4  Spider-Man: No Way Home   -2021
...
1670        Supergirl (2015-2021)
1681          Arrow  (2012-2020)
1682         Krypton (2018-2019)
1684     Black Lightning (2017-2021)
1685 DC's Legends of Tomorrow  (2016- )

          Genre  RunTime  IMDB_Score  IMDB_Score_scaled
0  Action,Adventure,Drama       0       0.0       0.000000
1  Action,Adventure,Fantasy     0       0.0       0.000000
2  Action,Adventure,Drama     50       7.5       0.781250
3  Action,Comedy,Drama     350       8.1       0.843750
4  Action,Adventure,Sci-Fi     0       0.0       0.000000
...
1670  Action,Adventure,Drama    42       8.6       0.895833
1681  Action,Adventure,Crime    42       8.7       0.906250
1682  Action,Adventure,Drama    41       7.8       0.812500
1684  Action,Drama,Sci-Fi     42       6.9       0.718750
1685  Action,Adventure,Drama    42       8.5       0.885417
```

[980 rows x 6 columns]

Melakukan normalisasi data pada kolom 'IMDB_Score' setelah proses deteksi outlier. Normalisasi ini bertujuan untuk mengubah skala nilai dalam rentang 0 hingga 1, sehingga memudahkan perbandingan dan analisis lebih lanjut. Setelah dinormalisasi, nilai dalam kolom 'IMDB_Score' menjadi lebih mudah dibandingkan karena skalanya sama.

```
1 data = pd.DataFrame(data)
2 scaler = MinMaxScaler()
3 data['RunTime_scaled'] = scaler.fit_transform(data[['RunTime']])
4 print(data)
```

```
          Movie      Year  \
0           Eternals    -2021
1             Loki   (2021- )
2 The Falcon and the Winter Soldier    -2021
3            WandaVision    -2021
4       Spider-Man: No Way Home    -2021
...           ...
1670        Supergirl (2015-2021)
1681          Arrow (2012-2020)
1682         Krypton (2018-2019)
1684        Black Lightning (2017-2021)
1685 DC's Legends of Tomorrow     (2016- )

          Genre RunTime  IMDB_Score  IMDB_Score_scaled  \
0  Action, Adventure, Drama      0        0.0        0.000000
1  Action, Adventure, Fantasy    0        0.0        0.000000
2  Action, Adventure, Drama     50        7.5        0.781250
3  Action, Comedy, Drama    350        8.1        0.843750
4  Action, Adventure, Sci-Fi     0        0.0        0.000000
...           ...
1670  Action, Adventure, Drama    42        8.6        0.895833
1681  Action, Adventure, Crime    42        8.7        0.906250
1682  Action, Adventure, Drama    41        7.8        0.812500
1684  Action, Drama, Sci-Fi     42        6.9        0.718750
1685  Action, Adventure, Drama    42        8.5        0.885417
```

```
          RunTime_scaled
0        0.000000
1        0.000000
2        0.094877
3        0.664137
4        0.000000
...
1670     0.079696
1681     0.079696
1682     0.077799
1684     0.079696
1685     0.079696
[980 rows x 7 columns]
```

Melakukan normalisasi data pada kolom 'RunTime' setelah proses deteksi outlier. Normalisasi ini bertujuan untuk mengubah skala nilai dalam rentang 0 hingga 1, sehingga memudahkan perbandingan dan analisis lebih lanjut. Setelah dinormalisasi, nilai dalam kolom 'RunTime' menjadi lebih mudah dibandingkan karena skalanya sama.

	Movie	Year	Genre	RunTime	IMDB_Score	IMDB_Score_scaled
0	Eternals	-2021	Action,Adventure,Drama	0	0.0	0.000000
1	Loki	(2021–)	Action,Adventure,Fantasy	0	0.0	0.000000
2	The Falcon and the Winter Soldier	-2021	Action,Adventure,Drama	50	7.5	0.781250
3	WandaVision	-2021	Action,Comedy,Drama	350	8.1	0.843750
4	Spider-Man: No Way Home	-2021	Action,Adventure,Sci-Fi	0	0.0	0.000000
5	Black Widow	-2021	Action,Adventure,Sci-Fi	133	0.0	0.000000
6	Avengers: Endgame	-2019	Action,Adventure,Drama	181	8.4	0.875000
7	Guardians of the Galaxy	-2014	Action,Adventure,Comedy	121	8.0	0.833333
8	Thor: Love and Thunder	-2022	Action,Adventure,Fantasy	0	0.0	0.000000
9	Spider-Man: Far from Home	-2019	Action,Adventure,Sci-Fi	129	7.5	0.781250

- Nilai IMDB Score yang mendekati 0 menunjukkan film yang kurang populer atau memiliki rating yang rendah, sedangkan nilai mendekati 1 menunjukkan film yang sangat populer atau memiliki rating yang tinggi.
- Dengan normalisasi, kita dapat membandingkan popularitas film secara lebih akurat, tanpa terpengaruh oleh perbedaan skala nilai asli.

FINAL DATA SET

Dataset terdiri dari 1685 baris dan 6 kolom

	Movie	Year	Genre	RunTime	IMDB_Score	IMDB_Score_scaled
1651	Watchmen	-2019	Action,Drama,Mystery	61	9.2	0.958333
1656	Watchmen	-2019	Action,Drama,Mystery	67	8.8	0.916667
1659	Black Lightning	(2017–2021)	Action,Drama,Sci-Fi	43	6.4	0.666667
1660	Black Lightning	(2017–2021)	Action,Drama,Sci-Fi	41	6.5	0.677083
1661	Black Lightning	(2017–2021)	Action,Drama,Sci-Fi	42	7.0	0.729167
1670	Supergirl	(2015–2021)	Action,Adventure,Drama	42	8.6	0.895833
1681	Arrow	(2012–2020)	Action,Adventure,Crime	42	8.7	0.906250
1682	Krypton	(2018–2019)	Action,Adventure,Drama	41	7.8	0.812500
1684	Black Lightning	(2017–2021)	Action,Drama,Sci-Fi	42	6.9	0.718750
1685	DC's Legends of Tomorrow	(2016–)	Action,Adventure,Drama	42	8.5	0.885417

GOOGLE COLAB

<https://colab.research.google.com/drive/1Wf3TvPliwsdTNqHs9aT-xtwOuc-7qMxv?usp=sharing>